一种综合事件本体相似度计算方法 *

朱文跃，刘　炜，刘宗田

(上海大学 计算机工程与科学学院, 上海 200444)

摘　要：事件本体相比于传统本体具有更加丰富的语义信息，在面向事件的大数据集成中更具优势，然而用传统的本体相似计算方法计算事件本体相似度存在很多不足，提出了一种综合的事件本体相似度计算方法。该方法以词语相似度、集合相似度、层次结构相似计算为基础，然后从事件类名称、事件类要素、事件类层次结构和非层次结构讨论事件本体的相似度，最终获得事件本体的综合相似度。实验表明该方法相比传统本体相似度计算方法准确率更高，语义信息更加丰富。

关键词：本体；事件本体；概念相似度；语义；相似度计算；事件本体相似度

中图分类号：TP391　　doi: 10.3969/j.issn.1001-3695.2018.01.0077

Comprehensive approach for event ontology similarity computation

Zhu Wenyue, Liu Wei, Liu Zongtian

(*School of Computer Engineering & Science*, *Shanghai University*, *Shanghai* 200444, *China*)

Abstract: Event ontology had more rich semantic information than traditional ontology, and had more advantages in event-oriented big data integration. There were many deficiencies in calculating the similarity of event ontology using traditional ontology similarity computation methods. So a comprehensive approach for calculating similarity of event ontology was proposed. This approach was based on words similarity computation, set similarity computation and hierarchical similarity computation, then, discussed event ontology similarity from the event class name, event class elements, event class hierarchy and event class non-hierarchy structure, finally got the comprehensive similarity of event ontology. Experimental results show that this approach is more accurate than the traditional ontology similarity computation approach and its semantic information is more abundant.

Key words: ontology; event ontology; concept similarity; semantic; similarity computation; event ontology similarity

0　引言

本体相似度计算是本体映射、本体集成、本体合并和本体翻译等技术的基础[1,2]，近年来本体相似度计算成为国内外知识计算领域（如本体对齐、数据集成）的研究热点。现有本体相似度计算方法主要分为基于元素的[3]、基于结构的[4]、基于实例的[5]和基于多策略集成的[6~8]几个部分。Abdul- Ghafour 等人[9]提出基于概念层次结构以及概念的属性进行语义相似度计算。近年来，随着基于自然语言的网络数据急剧增长，其数据类型多以叙事文本形式存在，利用传统本体对这类数据进行语义处理（如语义集成）存在很多不足[10]，利用传统本体很难刻画一个事件的完整信息（如什么时候、什么地点、发生可什么事，有哪些对象参加，事件发生的前后状态又是哪些影响）。然而事件本体是一种以事件类为基本单元的知识表示模型，相比于传统的以 “概念”为核心的本体，它可以描述包含事件动作、时间、对象、地点等要素的完整事件信息，可以保留更加丰富的语义内涵，更符合人类认知规律的知识[11]。因此，以事件为知识表示单元的事件本体模型近年来受到学术界的广泛关注[12]，事件本体的应用也逐步被研究人员所重视[13,14]。利用事件本体实现叙事类异构数据的语义集成是事件本体的一个重要应用方向，如利用事件本体将地震类新闻数据和台风灾害类新闻数据集成到自然灾害数据集中。基于事件本体的数据集成首先要解决事件本体的映射问题，即事件类与事件类之间的相似度计算问题。然而传统本体相似计算只是将事件类作为一类特殊的概念来处理，存在很多问题：a)概念离散问题，没有把事件类和事件类的参与者、时间、地点作为一个有机的整体来考虑；b)传

收稿日期：2018-01-31；**修回日期：**2018-03-23　　**基金项目：**国家自然科学基金资助项目（61273328，613050553）；上海市软科学研究计划资助项目（15692110200）

作者简介：朱文跃（1991-），男，安徽黄山人，硕士研究生，主要研究方向为主要研究方向为知识表示、知识的推理、机器学习等（zhuwenyue@t.shu.edu.cn）；刘炜（1978-），男，江西赣州人，副研究员，博士，主要研究方向为知识表示与推理、语义网与本体技术等；刘宗田（1946-），男，山东莒南人，研究员，博导，博士，主要研究方向为人工智能、软件工程、数据挖掘、粗糙集合、概念格等.

统本体相似度计算过程中，概念是单一的，只考虑了 “概念”层次结构来计算相似度，没有考虑到该事件类和其他事件类之间的关系；c)传统本体是基于静态“概念”的，很难刻画事件的动态性，没有考虑到时间、状态等因素的影响；d)事件本体模型对事件类有清晰的描述（如语言表现、描述事件类的核心词以及核心词搭配），这里其实包含了丰富的语义信息，传统本体方法计算相似度时往往都是忽略这些重要信息的。然而目前国内外对事件的相似度计算研究很少，因此，需要一种新的针对“事件”的本体相似度计算方法。基于事件本体相似度计算比传统的基于概念的本体相似度计算更加严谨准确，传统的本体相似计算只是考虑到了概念名称与概念之间的层次关系[15]，而事件本体相似度把事件类的参与者、时间、地点等要素作为一个有机的整体来计算，除了考虑到事件类与事件类之间的层次关系，还考虑到了事件类之间的非层次关系[12]，不仅如此，对于一个事件类的语言描述也囊括进了相似度计算，准确度更高，更能体现事件类之间的语义关系。

本文基于前期对事件本体模型的研究[11]，事件本体的相似度从事件类名称、事件类要素、事件类的层次结构和非层次结构四个方面考虑。事件类名称相似然后形成一个综合的事件本体相似度计算模型。其中事件类名称相似度包含名称的语法相似度和语义相似度，语义相似度计算借助于《知网》语义相似度来计算[18]。该模型不仅考虑事件类名称之间的语义信息，事件类各个要素之间的相似度，还考虑到构建事件本体时层次结构和非层次结构等信息，相比于传统基于概念的本体相似计算准确度更高。

1　事件本体相关概念

根据 Studder 等人[16]给出的定义，“本体是共享概念模型的明确的形式化规范说明”，本体的核心是概念与概念之间的关系。事件类 EC(event class)通常只是被作为一类特殊的概念来处理，很难表达出事件的本质，也很难描述事件类之间的复杂关系。事件本体是在传统本体基础上增加了对事件类及其关系的描述[10]。在事件不仅包含对象、时间、地点、动作等相关知识，而且包含了状态断言、语言表现等动态特征。因此以“事件”为中心的本体，可以很好地表达事件中的对象、时间、地点、动作以及事件和事件的一些复杂关系。

1.1　事件及事件类的定义

定义 1　事件(event)。某个特定时间地点下发生的，由一些角色参与的，表现出一些动作特征的一件事情。形式上用 e 来表示事件，事件由六要素组成，可以用六元组来表示[3]：

$$e ::= < A, O, T, P, S, L > \tag{1}$$

其中：A 表示动作，在文本中对应触发词，如“11 月 26 日 13 名恐怖分子在叙利亚被炸死”中的“炸死”；O 表示对象集合，表示参与事件的所有对象，上述例子中的“13 名恐怖分子”；T 表示时间，上述例子中的“11 月 26 日”；P 表示地点，上述例子中的“叙利亚”；S 表示状态集合，上述例子中“恐怖分子死了”；L 是语言表现，主要包括核心词集合和核心词搭配等。

定义 2　事件类(event class)。具有共同特征的事件集合[3]。用 EC 来表示定义如下：

$$EC = \{E, C_A, C_O, C_T, C_P, C_S, C_L\}$$

$$E = \{e_1, e_2, \ldots e_m\} (m \ge 0)$$

$$C_i = \{c_{i1}, c_{i2}, \cdots, c_{in}\} \quad (i \in \{A, O, T, P, S, L\}, n \ge 0) \tag{2}$$

其中：E 表示事件的集合，称之为事件类的外延；C_i 表示 E 中每个事件在第 i 个要素上具有的共同特征集合，称为事件类的内涵；C_{im} 表示 E 中每个事件在第 i 个要素上具有的一个共同属性。

OWL (Web ontology language)语言是 W3C 推荐的语义本体描述语言，用 OWL 对事件(类)要素进行扩展，将事件与事件要素之间的关系作为 ObjectProperty 对象属性来构建，如表 1 所示。在 OWL 中事件(类)与事件(类)要素之间的关系。如图 1 所示。

表 1　扩展的 OWL 中事件与事件要素之间的关系

要素名	Object	Time	Place	Action	Status	Language
关系名	Has Object	At Time	At Palce	Has Action	Has Status	Has Language

1.2　事件关系和事件本体结构

定义 3　事件类层次关系(hierarchy of event class) [13] $EC_1 = \{E_1, C_{1A}, C_{1O}, C_{1T}, C_{1P}, C_{1S}, C_{1L}\}$ 和事件类 $EC_2 = \{E_2, C_{2A}, C_{2O}, C_{2T}, C_{2P}, C_{2S}, C_{2L}\}$ EC_1 和 EC_2 存在分类关系，当且仅当（$E_1 \subset E_2$ 或者 $E_{1`} = E_2$ 且 $C_{1j} \subset C_{2j}$（$j \in \{A,O,T,V,P,L\}$））EC_1 称为 EC_2 的下位事件，EC_2 称为 EC_1 的上位事件。用 $R_{\text{is-a}}(EC_1, EC_2)$表示，如 “地震”和“交通事故”是“突发事件”的下位事件类。可以表为 $R_{\text{is-a}}$(地震，突发事件)，$R_{\text{is-a}}$(交通事故，突发事件)

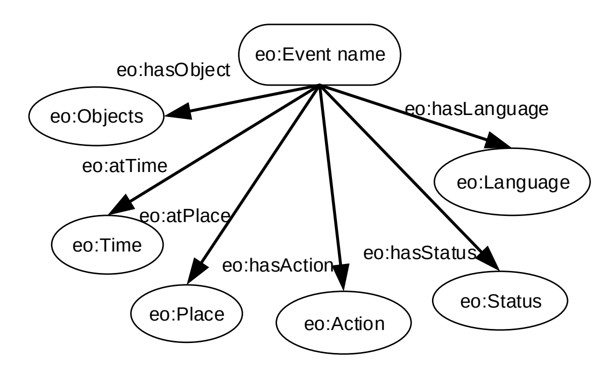


图 1　OWL 中一个事件的表示

定义 4　事件类非层次关系(non-hierarchy of event class)[12]，即：

a)组成关系。事件类 EC_1 是由事件类 EC_2 组成时，称两个事件类有组成关系，如“做饭”由“洗菜”和“烹饪”组成，可以表示为 $R_{\text{composedOf}}$(做饭，洗菜)、$R_{\text{composedOf}}$(做饭，烹饪)。

b)因果关系。事件类 EC_1 的发生以一定的概率导致事件类 EC_2 的发生。此时概率大于某个值时，称为两个事件类有因果关系，如“恐怖袭击”导致“平民伤亡”，R_{cause}(恐怖袭击，平民伤亡)。

c)跟随关系。在一定时间范围内，事件类 EC_1 的发生，以一定的概率跟随着事件类 EC_2 的发生，此时概率大于某个阈值时，称为两个事件类具有跟随关系，如“闪电”跟随着“打雷”，可以表示为 R_{follow}(闪电，打雷)。

d)并发关系。在一定时间范围内，事件类 EC_1 的发生，事件类 EC_2 以一定的概率同时发生，概率大于某个值时，称为两个事件类具有并发关系，如“刮风”和“下雨”可以表示为 R_{concur}(刮风，下雨)。

定义 5　事件本体(event ontology) 事件本体 EO 是共享客观存在的事件类模型[11]。其逻辑结构可定义为一个四元组:EO =< $UECS$， ECS， R， $Rules$， $Individuals$ >，其中:a) $UECS$ 顶层事件分类(Upper Event Class)集合；b) ECS 是事件类的集合 $ECS=\{EC_1, EC_2, \ldots EC_n\}$；c) $R=\{r\mid r$ 是事件(类)和事件(类)之间的关系，$r\in(R_{is\text{-}a}, R_{cause}, R_{composedOf}, R_{concur}, R_{follow})\}$；d) $Rules$ 是用逻辑语言表示的规则集合，包括事件类分类关系推理规则和事件关系推理规则；e) $Individuals$ 事件实例集合。事件本体模型结构如图 2 所示。上层事件类是通用的分类结构，下层事件类是通过事件关系构成的事件类格结构。

2　事件本体相似度计算方法

事件本体相似度计算主要分为三个部分介绍：第一部分，介绍词语的语法和语义相似度、集合的相似度、词语序列相似度和层次结构相似度；第二部分，根据上述相似度计算方法，对事件类名称、事件类要素、事件类层次结构和事件类非层次结构分别进行相似度计算；第三部分，将上述四个方面的相似度一起考虑进来进行综合相似度计算。

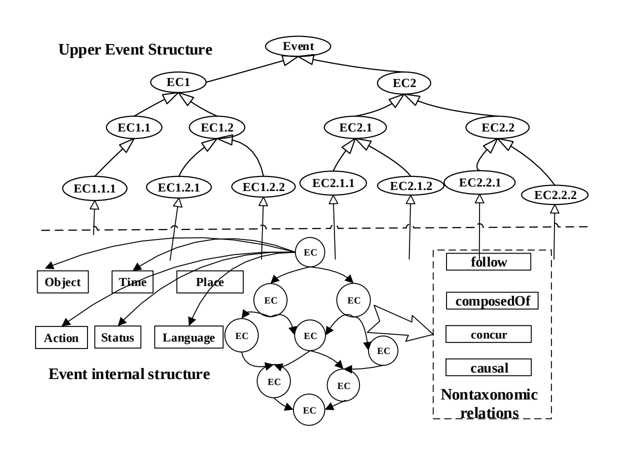


图 2　事件本体模型结构

2.1　相似度计算相关概念

2.1.1 语法相似度计算

对事件类名称计算方法有很多，如 Levenshtein、N-gram、Humming Distance 等方法。本文采用 Levenshtein 编辑距离[17]来计算两个字符串的相似度。两个字符串的相似度定义为

$$Sim_{syntax}(A,B)=\max\left(0,\frac{\min(|A|,|B|)-ed(A,B)}{\min(|A|,|B|)}\right) \tag{3}$$

其中：|A|、|B|分别是字符串 A、B 的长度；$\min(|A|, |B|)$表示是 A 与 B 中较短的字符串长度；$ed(A,B)$ 表示将 A 转化为 B 所需的最小操作数（包括插入、删除、替换等）。

2.1.2 语义相似度计算

《知网》(HowNet)作为语义资源基础，揭示概念和概念以及概念和概念所具有的属性之间的关系。因此可以将“事件”转换为“概念”进行相似度来计算。基于《知网》的词汇语义相似度计算作出了详细的阐述[18]。

1）义原的相似度计算

知网中的概念是通过义原来描述的，义原是描述概念的最小单位。两个义原节点之间的语义距离为 $Sim(p_1,p_2)$ 计算公式如下:

$$Sim(p_1,p_2)=\frac{\alpha}{d+\alpha} \tag{4}$$

其中：p_1、p_2 表示两个义原；d 表示 p_1、p_2 义原在层次体系中的路径长度；α 是一个调节因子。

2）概念的相似度计算

$$sim(S_1,S_2)=\sum_{i=1}^{4}\beta_i Sim_i(S_1,S_2) \tag{5}$$

其中：$\beta_i(1\le i\le 4)$ 是调节参数，并且 $\beta_1+\beta_2+\beta_3+\beta_4=1$，$\beta_1\ge\beta_2\ge\beta_3\ge\beta_4$；$Sim_1(S_1,S_2)$ 表示第一义原描述，$Sim_2(S_1,S_2)$ 表示其他基本义原描述，$Sim_3(S_1,S_2)$ 表示义原关系描述，$Sim_4(S_1,S_2)$ 表 θ_2 示关系符号描述。

3）词语相似度计算

对于两个汉语词语 W_1 和 W_2，假设在知网中 W_1 有 n 个义项(概念)[17]：$S_{11},S_{12},\ldots S_{1n}$，$W_2$ 有 m 个义项，词语 W_1 和词语 W_2 的相似度是其各个义项相似度的最大值。

$$Sim(W_1,W_2)=\max_{i=1\ldots n,\,j=1\ldots m} Sim(S_{1i},S_{2j}) \tag{6}$$

2.1.3 集合相似度计算

假设有集合 Set_1 和集合 Set_2，则集合的相似度为

$$t=\min(|set_1|,|set_2|)$$

$$Sim(Set_1,Set_2)=\frac{\sum_{i=1}^{t}Sim(W_{1i},W_{2i})}{t} \tag{7}$$

其中：$|Set_1|$和$|Set_2|$分别表示 Set_1 和 Set_2 中元素的个数；$\min(|set_1|,|set_2|)$ 表示 $|Set_1|$ 和 $|Set_2|$ 中较小的一个值；

$Sim(W_{1i},W_{2i})$ 的计算如下。将集合 Set_1 中所有元素和 Set_2 中的所有元素进行相似度计算得到的最大值作为 $Sim(W_{11},W_{21})$，将计算得到最大值对应的元素分别从 Set_1 和 Set_2 中删除，把 Set_1 剩余的元素看做新的 Set_1， Set_2 剩余的元素看做新的 Set_2，将新的集合 Set_1 中所有元素和新的集合 Set_2 中的所有元素进行相似度计算得到的最大值作为 $Sim(W_{21},W_{22})$。重复上述步骤得到 $Sim(W_{31},W_{32})$…，直到 Set_1 或者 Set_2 没有元素为止。

2.1.4 词语序列似度计算

对于事件类来说可能是用词语来描述的，如“地震”，也可能是短句描述的，如“交通事故事件”，因此需要对短句进行相似度计算。对短句进行相似度计算时，首先将短句进行分词保留实词部分，得到一个词语序列。假设两个短句得到的词语序列分别为 $Seq_1=(W_{11},W_{12},...,W_{1n})$ 和 $Seq_2=\{W_{21},W_{22},...W_{2n}\}$，则词语序列的相似度计算可以看做两个词语的集合，然后用式(7)进行计算。

2.1.5 层次结构相似度计算

事件本体中，无论是事件要素或者事件类都有可能是层次结构，因此，在这里给出层次结构相似度计算方法。层次结构相似度算法很多，如 Resnik 算法[19]，定义两个节点的相似度为其最低共同祖先节点的信息量。文献[20]在最低共同祖先节点的基础上增加了共享路径的层次结构相似度计算。下面介绍本文的层次结构相似度计算算法。节点深度是指节点在层次结构中所处的层数，用 $depth(N)$ 表示节点 N 的深度，根节点深度为 1。层次结构相似度从下面三个方面来考虑：

a)两个节点深度总和，即 $depth(N_A)+depth(N_B)$；在路径距离相同情况下，节点深度总和越大，相似度也越大。

b)两个节点的最近公共父节点深度 $depth(N_P(N_A,\ N_B))$，其中，$N_P(N_A,N_B)$ 表示节点 N_A 和 N_B 的最近公共父节点，最近公共父节点越深，它的分类就越细致，继承的信息量也就越多，则相似度越大。

c)两个节点的相对深度，即两个节点深度绝对值之差 $|depth(N_A)-depth(N_B)|$； $S_{21},S_{22},...S_{2m}$ 相对深度越小层次差异越小，相似度越大。基于上述思想定义两个节点的相似度：

$$Sim_{structure}(N_A-N_B)=\theta_1\alpha+\theta_2\beta+\theta_3\lambda \tag{8}$$

$$\lambda=\frac{\max depth}{\max depth+|depth(N_A)-depth(N_B)|} \quad \alpha=\frac{2\times depth(N_P(N_A,N_B))}{depth(N_A)+depth(N_B)}$$

$\beta=\dfrac{depth(N_P(N_A,N_B))}{\max depth}$， $\max depth$ 表示事件本体层次结构的最大深度， θ_1、θ_2、θ_3 为权值 $\theta_1+\theta_2+\theta_3=1$。本文层次结构相似度计算过程中，权值 θ_1 为 0.4、θ_2 为 0.3、θ_3 为 0.3。

2.2 事件名称相似度计算

事件类名称相似度计算，需要从语法和语义两个方面来考虑。假设两个事件类名称分别为 name1、name2，事件类名称语法相似度 $Sim_{syntax}(name1,name2)$ 可以用式(3)来计算。

事件类名称语义相似度 $Sim_{semantic}(name1,name2)$ 计算时，如果两个事件类都是词语，可以用词语相似度计算，见式(6)。如果两个事件类中有短句，如“醉酒驾驶”先对短句进行分词保留实词，得到一个词语序列，然后通过词语序列相似度计算来求得事件类名称的语义相似度，最后结合语法相似度和语义相似得出事件类名称综合相似度 $Sim(name1,name2)$。

$$Sim(name1,name2)=0.3\times Sim_{syntax}(name1,name2)+0.7\times Sim_{semantic}(name1,name2)) \tag{9}$$

2.3 事件类要素相似度计算

事件类和事件一样都是由六个要素构成，事件类的状态要素描述的是事件发生的状态变化。暂时不考虑状态要素。在事件类六要素中，每个要素都有其独特的特征。因此计算事件类要素相似度需要分开考虑。

2.3.1 事件类动作要素相似度计算

事件类中的动作要素是事件类的触发词（指示词）[12]，见表 2。一般认为只要触发词属于同一事件类型，动作要素就相似度为 1，如地震、余震和震感等动作要素相似度都为 1。如果不是属于同一事件类的触发词，则将触发词看做词语，每个事件类的触发词构成词语集合，采用式(7)来计算两个事件类的动作要素相似度。

表 2 事件类的触发词

事件类型	触发词	事件类型	触发词
地震	地震、余震、震感…	食物中毒	中毒、呕吐、恶心…
交通事故	追尾、撞车…	火灾	着火、燃烧…
损失	倒塌、烧毁、损坏…	伤亡	死亡、丧生、受伤…

2.3.2 事件类时间要素相似度计算

时间要素相似度计算可以借助于传统的时间本体（time onotology）计算其相似度，也可以根据需求构建层次结构进行相似度计算。本文借助传统时间本体[21,22]，通过对时间的描述匹配判断两个事件类时间要素是否相似，如图 3 所示。例如，历史事件类，一般表述为公元某年发生了某事件，是以时间方向来描述的。然而新闻报道类事件，一般表述为北京时间几点几分发生了某事件，以时区/标准时间来报道。如果两个事件类都有相同的时间格式，则认为相似，否则认为不相似。对于一般事件类不存在时间要素的话，就不必考虑时间要素相似度。

2.3.3 事件类地点要素相似度计算

事件类的地点要素一般为地点结合事件类的层次结构相似度计算。根据国土资源部在 2010 年发布的《县级土地利用总体规划编制规程》的“土地规划分类及其含义”，提取地点要素实体、概念，构建地点要素本体层次结构，见图 4。地点要素是层

次结构，相似度计算按照式(8)。

2.3.4 事件类对象要素相似度计算

事件类中的对象要素，是参与到事件中的对象集合，包括事件的施动者、受动者、工具等。两个事件类的对象要素是两个集合，可以把对象看做词语，最后构成的是两个词语集合，可以通过式(7)来计算对象要素的相似度。

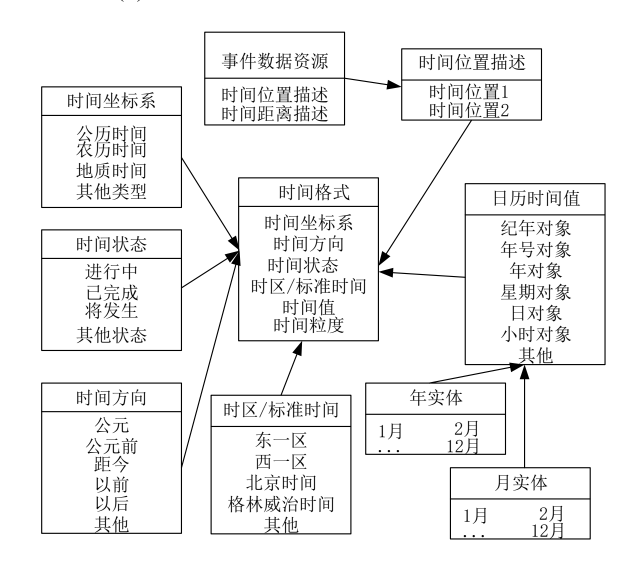


图 3 时间要素的描述

2.3.5 事件类语言表现要素相似度计算

事件类中的语言表现要素，主要体现在两个方面：核心词和核心词搭配。本文暂且不考虑核心词搭配问题。语言表现中的核心词构成了词语集合，可以通过式(7)相似度计算公式来计算事件类语言要素相似度。

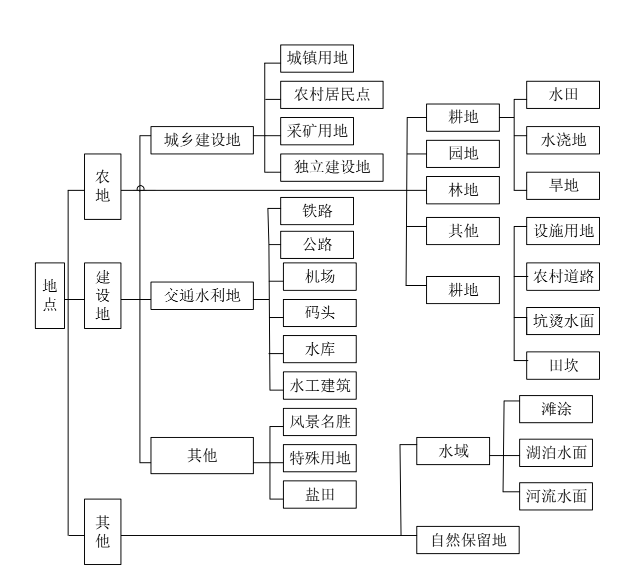


图 4 地点要素本体局部层次网络结构

2.3.6 事件类要素综合相似度计算

假设事件类 EC_1 和 EC_2，两个事件类要素相似度为综合相似度为 $Sim(EC_1,EC_2)$，计算公式如下:

$$Sim_{\text{element}}(EC_1,EC_2)=\sum_{i=1}^{5}\alpha_i Sim(\text{element}_1,\text{element}_2)$$

$$(\text{i}=\text{A,O,T,P,L})\ \sum_{i=1}^{5}\alpha_i=1 \tag{10}$$

其中：$element_1$ 表示事件类 EC_1 的事件类要素；$element_2$ 表示为事件类 EC_2 的事件类要素；α_i 取值在 0~1 间，表示事件各要素的权重。本文为了方便计算，认为要素相似度重要程度相同，取值相等都为 0.2.

2.4 事件类层次结构相似度计算

在事件类关系有分类关系..和非分类关系(R_{cause} ， $R_{compositedOf}$ ， $R_{concurrence}$ ， R_{follow})。对于 R_{is-a} 关系，即层次结构关系。把事件类看做层次结构中的节点，采用式（8）计算事件类层次结构相似度。

2.5 事件类非层次结构相似度计算

事件（类）之间非层次关系是指通过(R_{cause} ， $R_{compositedOf}$ ， $R_{concurrence}$ ， R_{follow})四种关系相连。假设每种关系的重要程度相同。本文取一个语义半径 R，事件类 EC_A 为中心，以路径长度 P 为半径，使得 $P\le R$ 来选取周围事件类。假设得到事件类集合为 $ECS_A=\{\text{EC}_{A1},\text{EC}_{A2},...\text{EC}_{Am}\}$。事件类 EC_B 为中心，以路径长度 P 为半径，使得 $P\le R$ 来选取周围事件类。假设得到事件类集合为 $ECS_B=\{\text{EC}_{B1},\text{EC}_{B2},...\text{EC}_{Bn}\}$，那么两个事件类的非层次结构相似度 $Sim(\text{EC}_A,\text{EC}_B)$ 为

$$Sim_{nonstru}(\text{EC}_\text{A},\text{EC}_\text{B})=\frac{2\times|\text{count}(\text{ECS}_\text{A}\cap\text{ECS}_\text{B})|}{|\text{count}(\text{ECS}_\text{A})|+|\text{count}(\text{ECS}_\text{B})|} \tag{11}$$

其中：$|\text{count}(\text{ECS}_\text{A}\cap\text{ECS}_\text{B})|$ 表示既在 ECS_A 又在 ECS_B 中的事件类的个数；$|count(\text{ECS}_\text{A})|$ 表示 ECS_A 中事件类个数；$|count(\text{ECS}_B)|$ 表示 ECS_B 中事件类个数。本文计算非层次结构相似度中语义半径 R 取值为 1。

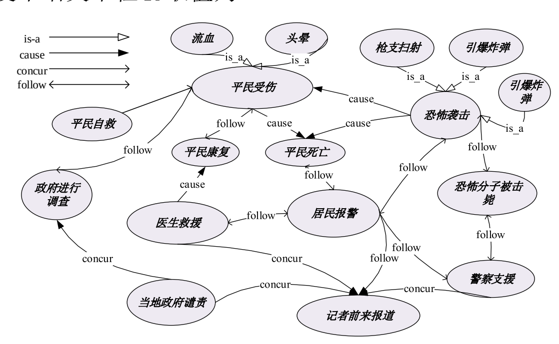


图 5 恐怖袭击事件类非层次结构

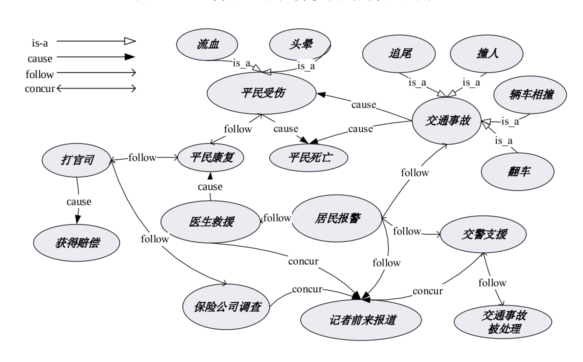


图 6 交通事故事件类非层次结构

2.6 事件类综合相似度计算

基于上述分析，综合考虑基于事件类名称的相似度、基于事件类要素的相似度、基于层次结构的相似度和基于非层次结构的相似度，最后得到综合相似度计算公式如下：

$$Sim(\mathrm{EC}_A,\mathrm{EC}_B)=w_1Sim_{name}(\mathrm{EC}_A,\mathrm{EC}_B)+w_2Sim_{element}(\mathrm{EC}_A,\mathrm{EC}_B)+$$
$$w_3Sim_{structure}(\mathrm{EC}_A,\mathrm{EC}_B)+w_4Sim_{nstru}(\mathrm{EC}_A,\mathrm{EC}_B)\quad \sum_{i=1}^{4}w_i=1 \tag{12}$$

其中：w_i 表示权值；EC_A 和 EC_B 表示事件类；$Sim_{name}(\mathrm{EC}_A,\mathrm{EC}_B)$ 表示 EC_A 、与 EC_B 名称相似度；$Sim_{element}(\mathrm{EC}_A,\mathrm{EC}_B)$ 表示 EC_A 、与 EC_B 要素综合相似度；$Sim_{structure}(\mathrm{EC}_A,\mathrm{EC}_B)$ 表示 EC_A 、与 EC_B 结构相似度；$Sim_{nstru}(\mathrm{EC}_A,\mathrm{EC}_B)$ 表示 EC_A、与 EC_B 非结构相似度。名称相似度权重 w_1 为 0.3，事件类要素相似度权重 w_2 为 0.2；事件类结构相似度权重 w_3 为 0.4，事件类非结构相似度权重 w_4 为 0.1。

3 案例分析和实验结果

以《国家突发公共事件总体应急预案》的事件分类体系为依据，参照《突发公共卫生事件流行病学》中突发事件的分类结构，结合 CEC（Chinese Emergency Corpus）中文突发事件语料库 332 篇语料[23]和新浪新闻网上爬取的 300 篇报道，构建相应的事件本体层次结构，如图 7 所示。把突发事件类别体系分为三个层次：第一层 4 大类，第二层 25 子类，第三层 80 小类，第四层 243 个事件类。

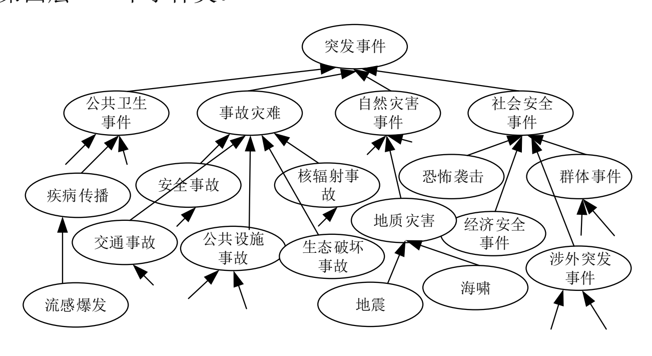


图 7 事件本体中事件类部分层次结构图

3.1 实例分析

以恐怖袭击事件类和交通事故事件类为例进行分析。两个事件类六要素见表 3。在领域专家的指导下，本文建立了恐怖袭击事件类的非层次结构，如图 5 所示，建立了交通事故事件类非层次结构，如图 6 所示。

表 3 交通事故和恐怖袭击事件类六要素

事件类	交通事故事件类	恐怖袭击事件类
时间要素	日历时间/时区时间	日历时间/时区时间
地点要素	公路	建设地
对象要素	车，人，红绿灯	人、炸弹、建筑物等
动作要素	追尾、撞车	爆炸、枪击、袭击
状态要素	前置状态：人车完好、道路交通良好 后置状态：人伤亡、车受损、交通拥堵	前置状态：人安全、建筑物完好、炸弹未爆炸 后置状态：人伤亡、建筑物毁坏、炸弹已爆
语言表现	核心词：碰撞、相撞、车祸、追尾、卧轨、刹车、闯红灯、受伤等 核心词搭配：发生碰撞 出现车祸、两车追尾	核心词：爆炸、枪击、绑架、劫持、纵火、平民受伤等 核心词搭配: 炸弹爆炸、发生枪击、劫持人质、恶意纵火

3.2 实验结果分析

在事件本体中，随机选取两个事件类来做相似度计算，按照式(12)、文献[20]和文献[24]算法进行比较，如表 4 所示。文献[24]算法是在 Resnik 算法的基础上增加了共享路径的层次结构相似度计算。该算法可根据不同本体的结构进行节点和有向边的权重分配，比 Resnik 算法结果更加准确一点。文献[24]通过分析义项的描述语言结构，将关系义原和关系符号描述结构进行加权然后进行相似计算，是一种改进的《知网》相似度计算方法。

表 4 事件类相似度计算实验数据

事件类 1	事件类 2	本方法	文献[20]	文献[24]
恐怖袭击	交通事故	0.421	0.322	0.522
事故灾难	交通事故	0.604	0.754	0.257
冰雹	洪水	0.352	0.420	0.187
泥石流	地震	0.500	0.627	0.200
地震	海啸	0.781	0.712	0.981
流感	地震	0.483	0.213	0.552

可以发现，本文算法与文献[20]计算出来的相似度大体上相似，如“恐怖袭击”与“交通事故” “事故灾难”与“群体事件”相似度的值很接近。这是因为两种算法，层次结构相似度在整个相似计算过程中的权重比较大。然而在“流感”和“地震”两个事件类相差比较大，因为文献[20]只考虑层次结构相似度，没有考虑非结构层次结构相似度。然而这两种算法“地震”事件类和“海啸”事件类相似度上差别比较大。因为文献[20]只是用到了层次结构信息，没有用到的《知网》中的语义信息，也没有用到事件类要素之间的信息，所以其相似度计算出来的值没有本算法高。

本文算法与文献[24]作比较，“地震”和“海啸”相似度差别比较大，因为《知网》中“地震”和“海啸”的义原集合以及义原层次结构都很相似，导致“地震”和“海啸”相似值接近于 1。本算法中除了考虑层次结构，还考虑了非层次结构，更加符合事件发生的规律。比如，“地震”发生一般会导致房屋倒塌、山体滑坡、发生燕塞湖等灾害，而海啸不会，两者之间还是有些区别的。另外文献[24]只是将“事件类”作为概念来处理，没有考虑到事件类的六要素，没有考虑到事件类之间的非层次关系，忽略了事件类丰富的语义信息。所示计算结果不是很准确。

从实验结果可以看出本算法结合了《知网》的语义特性，还考虑到了事件类六要素的信息以及事件类的非层次结构，相比于传统的本体相似度计算算法，考虑更加全面，计算结果也更为准确。

4 结束语

“事件”作为人类知识的单元，包含丰富的语义信息。本

文提出了事件本体相似度方法，从事件类名称、事件类六要素、事件类层次结构、事件类非层次结构讨论了事件类的相似度计算。本方法相似度计算时，不仅考虑到基于“概念”的相似度计算，还考虑到了，“事件类”丰富的语义信息。在传统本体层次结构相似度基础上，增加了“事件类”的非层次结构相似度计算。这使得计算结果更能全面准备地放映出“事件类”之间的相似度。然而，也有需要进一步改善的地方，首先实验过程中大量权重还是根据经验来确定的。另外，事件类要素相似度计算过程中需要详细标注的事件本体。因此，下一步将改进计算过程中各种权重设定的方法，并考虑在缺少详细标注的事件本体计算相似度。

参考文献：

[1] Stojanovic L. Methods and tools for ontology evolution [J]. Information Systems Methodology, 2004, 16 (1): 411-423.

[2] Sanchez D, Batet M, Isern D, *et al.* Ontology-based semantic similarity: a new feature-based approach [J]. Expert Systems with Applications, 2012, 39 (9): 7718-7728.

[3] 刘秀磊，廖建新，朱晓民，等. 本体匹配中基于词义组合的词法分析算法 [J]. 电子学报, 2012, 40 (8): 1624-1630. (Liu Xiulei, Liao Jianxin, Zhu Xiaomin, *et al.* Lexical analysis based on combining senses in ontology matching [J]. Acta Electronica Sinica, 2012, 40 (8): 1624-1630.)

[4] Hu W, Qu Y, Cheng G. Matching large ontologies: a divide-and-conquer approach [J]. Data and Knowledge Engineering, 2008, 67 (1): 140-160.

[5] Pirro G, Talia D. an ontology mapping system with strategy prediction capabilities [J]. Data and Knowledge Engineering, 2010, 69 (5): 444-471.

[6] Belhadefa H. A new bidirectional method for ontologies matching [J]. Procedia Engineering, 2011, 23 (23): 558-564.

[7] Vargas-Vera M, Ibanez U A, Mar Vd, *et al.* State of the Art on Ontology Alignment [J]. International Journal of Knowledge Society Research, 2015, 6 (1): 17-42

[8] 张忠平，田淑霞，刘洪强. 一种新的本体相似度计算方法 [J]. 计算机应用研究, 2008, 25 (10): 2929-2942. (Zhang Zhongping, Tian Shuxia, Liu Hongqiang. New approach for ontology similarity computation [J]. Application Research of Computers, 2008, 25 (10): 2929-2942.)

[9] Abdul-Ghafour S, Ghodous P, Shariat B, *et al.* Semantic interoperability of knowledge in feature-based CAD models [J] Computer Aided Design, 2014, 56 (11): 45-57.

[10] 刘宗田，黄美丽，周文，等. 面向事件的本体研究 [J]. 计算机科学, 2009, 36 (11): 189-192. (Liu Zongtian, Huang Meili, Zhou Wen, *et al.* Research on event-oriented ontology model [J]. Computer Science, 2009, 36 (11): 189-192.)

[11] 刘炜，丁宁，杨竣辉，等. 针对地点污染突发事件领域的事件本体模式 [J]. 计算机科学与探索, 2016, 10 (4): 446-480. (Liu Wei, Ding Ning, Yang Junhui, *et al.* Event ontology pattern for domain of environmental pollution emergency [J]. Journal of Frontiers of Computer Science and Technology, 2016, 10 (4): 446-480.)

[12] 仲兆满，刘宗田，李存华. 事件本体模型及事件类排序 [J]. 北京大学学报：自然科学版, 2013, 49 (2): 234-240. (Zhong Zhaoman, Liu Zongtian, Li Cunhua. Event ontology model and event class ranking [J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2013, 49 (2): 234-240.)

[13] Zhou Wen, Zhang Yajun, Su XiaoYing, *et al.* Semantic role labeling based event argument identification [J]. International Journal of Database Theory Application, 2016, 9 (6): 93-102.

[14] Zhang Yajun, Liu Zongtian, Zhang Yalan. Formal research of event ontology object elements and recognition of abbreviated object [J]. Journal of Computational Information Systems, 2015, 11 (22): 8039-8049.

[15] Montani S, Leonardi S, Quaglini S, *et al.* A knowledge-intensive approach to process similarity calculation [J]. Expert Systems with Applications, 2015, 42 (9): 4207-4215.

[16] Studer R, Benjamins VR, Fensel D. Knowledge engineering: principles and methods [J]. Data and knowledge engineering, 1998, 25 (1): 167-197.

[17] Levenshtein VI. Binary codes capable of correcting deletions, insertions, and reversals [J]. Soviet Physics Doklady, 1966, 10 (8): 707-710.

[18] 刘群，李素建. 基于《知网》的词汇语义相似度计算 [C]// 第三届汉语词汇语义学研讨会论文集. 2002: 59-76. (Liu Qun, Li Sujian. Word similarity computing based on how-net [C]// Proc of the3rd Chinese Lexical Semantics Seminar. 2002: 59-76.)

[19] Resnik P. Using information content to evaluate semantic similarity in a taxonomy [C]// Proc of the 14th International joint conference on Artificial Intelligence. San Francisco CA: Morgan Kaufmann Publishers Inc, 1995: 448-453.

[20] 甘明鑫，窦雪，王道平等. 一种综合加权的本体概念语义相似度计算方法 [J]. 计算机工程与应用, 2012, 48 (17): 148-153. (Gan Mingxin, Dou Xue, Wang Daoping, *et al.* Comprehensive weighting method for calculation of ontoIogy-based semantic simiIarity [J]. Computer Engineering and Applications, 2012, 48 (17): 148-153)

[21] Zhang C X. Cao C G. Sui Y F, *et al.* A Chinese time ontology for the Semantic Web [J]. Knowledge-Based Systems, 2011, 24 (7): 1057-1074.

[22] W3C. Time ontology in OWL [EB/OL]. [2018-03-06]. https://www. w3. org/TR/owl-time.

[23] 上海大学智能语义实验室. CEC 中文突发事件语料库 [EB/OL]. [2018-03-06]. https://github. com/daselab/CEC-Corpus. (Shanghai University. Chinese Environment Emergency Corpus [EB/OL]. [2018-03-06]. https://github. com/daselab/CEC-Corpus.)

[24] 葛斌，李芳芳，郭丝路，等. 基于知网的词汇语义相似度计算方法研究 [J]. 计算机应用研究, 2010, 27 (9): 3329-3333. (Ge Bin, Li Fangfang, Guo Silu, *et al.* Word’s semantic similarity computation method based on Hownet [J]. Application Research of Computers, 2010, 27 (9): 3329-3333.)